数学基础班第 3 课课件：概率论选讲

管枫

七月在线

July, 2016

主要内容

- 积分学
  - 理解积分：无穷求和，体积
  - 微积分基本定理：牛顿-莱布尼茨公式
- 概率空间
  - 随机变量与概率：概率密度函数的积分
  - 条件概率
  - 共轭分布
- 大数定律和中心极限定理
  - 随机变量的矩
  - 切比雪夫不等式
  - 大数定律
  - 中心极限定理

记号

- 本节课常用数学记号

$\mathbb{R}^n$	实坐标空间
$f(x)$	函数
$\int_a^b f(x)dx$	函数积分
X	随机变量
$E(X)$	随机变量的期望
$Var(x)$	随机变量的方差
$\mu_n(x)$	随机变量的 n 阶矩

Definition (单变量函数黎曼积分)

令 $f(x)$ 为开区间 (a,b) 上的一个连续函数，对于任何一个正整数 n 定义,$x_i = a + \frac{i(b-a)}{n}$ 求和式:

$$S_n(f) = \sum_{i=0}^{n-1} f(x_i)(x_{i+1} - x_i)$$

如果极限 $\lim_{n\to\infty} S_n(f)$ 存在, 那么函数 $f(x)$ 在这个区间上的黎曼积分为

$$\int_a^b f(x)dx = \lim_{n\to\infty} S_n(f)$$

理解积分

- 代数意义：无穷求和
- 几何意义：函数与 X 轴之间的有向面积

此处课堂画图举例说明

积分学: 微积分基本定理：牛顿-莱布尼茨公式

Theorem (牛顿-莱布尼茨公式)

如果 $f(x)$ 是定义在闭区间 $[a, b]$ 上的可微函数, 那么就有

$$\int_a^b f'(t)dt = f(b) - f(a)$$

不定积分表示为

$$\int f'(t)dt = f(x) + C$$

牛顿-莱布尼茨公式展示了微分与积分的基本关系: 在一定程度上微分与积分互为逆运算.

Example

函数 $\ln(x)$ 的不定积分

令 $f(x) = x\ln(x) - x$，则 $f'(x) = 1 \cdot \ln(x) + x \cdot \frac{1}{x} - 1 = \ln(x)$.
根据牛顿-莱布尼茨公式我们得到

$$\int \ln(t)dt = \int f'(t)dt = x\ln(x) - x + C$$

Example

函数 $\frac{1}{x^\alpha} = x^{-\alpha}$ 的不定积分

我们知道幂函数的导数公式 $\frac{d}{dx}x^n = n \cdot x^{n-1}$.
反向思维如果 $n-1=-\alpha$，那么 $n=-\alpha+1$. 所以令
$f(x) = x^{-\alpha+1}$，那么 $f'(x) = (-\alpha+1)\cdot x^{-\alpha}$.
于是当 $\alpha \neq 1$ 时，

$$\int \frac{1}{x^\alpha}dx = \frac{1}{(1-\alpha)x^{\alpha-1}} + C$$

当 $\alpha = 1$ 的时候呢，我们熟知 $\frac{d}{dx}\ln(x) = 1/x$. 所以

$$\int \frac{1}{x}dx = \ln(x) + C$$

多变量函数的积分

如果一个函数 $f(x,y)$ 有多个变量，那么在矩形 $[a,b]\times[c,d]$ 上的多重积分可以看成是每一个变量的依次积分

$$\int_c^d \int_a^b f(x,y)dxdy$$

- 如果积分区域形状不规则，可以用一个矩形把积分区域包起来，并令函数在积分区域外边等于 0.
- 二重积分的几何意义是积分函数与 $X-Y$ 坐标平面之间部分的有向体积.

小结 (积分学)

- 积分的代数意义是无穷求和，几何意义是带符号的体积
- 微分和积分在一定程度上互为逆运算
- 熟悉微分公式有助于计算积分
- 多重积分可以理解成是依次进行的单重积分

随机变量与概率：概率密度函数的积分

离散随机变量

假设随机变量 X 的取值域为 $\Omega = \{x_i\}_{i=1}^{\infty}$，那么对于任何一个 x_i，事件 $X = x_i$ 的概率记为 $P(x_i)$.

对于 Ω 的任何一个子集 $S = \{x_{k_i}\}_{i=1}^{\infty}$，事件 $X \in S$ 的概率为

$$P(S) = \sum_{i=1}^{\infty} P(x_i)$$

对于离散随机变量，概率为概率函数的求和.

连续随机变量

假设随机变量 X 的取值域为 $\mathbb{R}$，那么对于几乎所有 $x \in \mathbb{R}$, 事件 $X = x$ 的概率 $P(X = x)$ 都等于 0. 所以我们转而定义概率密度函数 $f : \mathbb{R} \to [0, \infty)$. 对于任何区间 (a, b), 事件 $X \in (a, b)$ 的概率为

$$P((a,b)) = \int_a^b f(x)dx$$

- 对于连续型随机变量，概率为概率密度函数的积分.
- 不论是离散还是连续型随机变量, 概率函数和概率密度函数的定义域即为这个随机变量的值域.
- 作为一个特殊的概率函数，分布函数定义为 $\Phi(x) = P(X < x)$.

我们在此课中只考虑几乎处处连续的概率密度函数，我们不考虑离散，连续混合型的随机变量

随机变量与概率：如何理解概率

事件的概率

- 整个概率空间是一个事件，这个事件一定发生所以全空间的概率为 1
- 事件是随机变量值域的子集 S
- 事件的概率则表示 S 里面概率之和或概率密度之积分.

事件的条件概率

- 条件也是事件，也可表示为随机变量值域的子集:A
- 条件概率里面的事件，又是这个条件的子集:$S \cap A \subset A$
- 事件的条件概率则表示 $S \cap A$ 在 A 里面所占的比例. 故而 $P(S|A) = \frac{P(S \cap A)}{P(A)}$

概率其实就是集合的大小比例，而概率函数或者概率密度函数可以理解为比较大小时候的权重

贝叶斯公式

如果 A, B 是两个事件，那么条件概率满足公式

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

利用前面的定义我们知道，事件 A, B 同时发生的概率为 $P(A \cap B)$, 一方面

$$P(A \cap B) = P(B|A)P(A)$$

另一方面对称的有

$$P(A \cap B) = P(A|B)P(B)$$

所以 $P(B|A)P(A) = P(A|B)P(B)$, 两边同时除以 $P(B)$ 就得到了贝叶斯公式.

随机变量与概率：共轭分布

常见的概率分布基本上都有参数，比如正态分布有 (μ, σ) 两个参数，泊松分布有一个参数 λ. 对于一个具体的问题而言，关于这些参数有两种不同的看法

- 利用经验得到一个关于参数的先验分布.(Bayesian)
- 不对参数先验分布做任何假设，只利用当前观测的数据来对参数进行估计.(Frequentist)

本课对这两种观点相关的方法都进行介绍，本章介绍的先验分布，以及共轭分布将来会出现在贝叶斯学习中，而下一章的极大似然估计方法就更像是 Frequentists 的做法

先验分布，似然函数，后验分布

- 参数先验分布为 $p(\theta)$
- 似然函数为 $p(x|\theta)$
- 观测值为 X

贝叶斯的思想是根据观测值来调整参数的先验分布从而得到参数的后验分布. 参数后验分布为

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_{\theta'} p(X|\theta')p(\theta')d\theta'}$$

共轭分布

如果参数的后验分布与先验分布属于同一类分布，那么我们说这种先验分布为共轭分布 (Conjugate prior). 比如

- 似然函数为正态分布时, 如果 σ 已知，关于 μ 的正太分布是共轭分布
- 似然函数为正态分布时, 如果 μ 已知，关于 σ 的反 Gamma 分布是共轭分布

具体共轭分布列表可以参考 https://en.wikipedia.org/wiki/Conjugate_prior

共轭分布的好处在于，先验与后验分布属于一个大类，这样计算和理解上都比较方便.

小结 (随机变量与概率)

- 概率可以理解为事件所代表的集合在全概率空间中的比例
- 对于概率分布参数的先验分布有不同的观点
- 如果参数先验分布与后验分布属于同一类，则叫做共轭分布

随机变量的矩

X 是一个随机变量对于任何正整数 n，定义

$$E(X^n) = \int p(x)x^n dx$$

- 当 $n = 1$ 时，$E(X)$ 为随机变量的期望
- 当 $n = 2$ 时，$E(X^2) - E(X)^2$ 为随机变量的方差
- 特征函数，$E(e^{itX}) = \sum_{n=0}^{\infty} \frac{E(X^n)}{n!}(it)^n$.

矩可以描述随机变量的一些特征，期望是 X “中心”位置的一种描述，方差可以描述 X 的分散程度, 特征函数可以全面描述概率分布.

切比雪夫不等式

设 X 为随机变量，期望值为 μ, 标准差为 σ, 对于任何实数 $k>0$

$$P(|X-\mu|\geq k\sigma)\leq\frac{1}{k^2}$$

切比雪夫不等式给出方差对 X 分散程度的描述提供了一个定量的估计.

$$
\begin{aligned}
P(|X-\mu| \geq k\sigma) &= \int\limits_{|X-\mu|\geq k\sigma} p(x)dx = \int\limits_{|X-\mu|\geq k\sigma} \frac{x^2}{x^2}p(x)dx \\
&\leq \int\limits_{|X-\mu|\geq k\sigma} \frac{x^2}{k^2\sigma^2}p(x)dx \\
&= \frac{1}{k^2\sigma^2}\int\limits_{|X-\mu|\geq k\sigma} x^2p(x)dx \\
&\leq \frac{1}{k^2\sigma^2}\int x^2p(x)dx \\
&= \frac{\sigma^2}{k^2\sigma^2} \\
&= \frac{1}{k^2}
\end{aligned}
$$

随机变量的相关系数

X,Y 是两个随机变量

- X, Y 的协方差：$cov(X,Y) = E(XY) - E(X)E(Y)$
- X, Y 的相关系数：$cov(X,Y)/\sqrt{(Var(X)Var(Y))}$

独立随机变量

X,Y 是两个随机变量如果联合分布 $p(x,y) = p(x)p(y)$, 则 X, Y 为独立随机变量.

- 独立随机变量相关系数为 0
- 相关系数为零，两个随机变量不见得独立

如下列举一些特征函数 $\phi_X(t) = E(e^{itX})$ 自身的重要性质

利用特征函数研究概率分布

- 对于任何 X,$\phi_X(t)$ 都存在
- $\phi(0) = 1$ 且 $|\phi(t)| \leq 1, \forall t$
- $\phi(t)$ 是一致连续函数
- $\overline{\phi_X(t)} = \phi_{-X}(t)$, 所以如果 X 关于中心对称，那么 $\phi_X(t)$ 就是一个实函数
- 如果 X 的 n 阶矩存在，那么 $\phi_X(t)$ 至少 n 阶可微，并且 $E(X^n) = (-i)^n\phi^{(n)}(0)$

如下列举一些不同的随机变量的特征函数的重要性质

利用特征函数研究概率分布

- 如果 X, Y 是两个独立随机变量，那么 $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$
- 如果 $\phi_X(t) = \phi_Y(t)$, 那么 X, Y 服从同一个分布.
- 如果 $\{X_n\}$ 是一个随机变量序列，而且 $\phi_{X_n}(t)$ 逐点收敛于一个函数 $\phi_\infty(t)$, 如果 ϕ_∞ 在 0 处连续，那么存在一个分布 $X_\infty(t)$, 使得 X_n 按分布收敛于 $X_\infty(t)$.

特殊分布的特征函数

- 独点分布 $p(a)=1$， $\phi(t)=e^{iat}$
- 两点分布 $p(-1)=p(1)=\frac{1}{2}$,$\phi(t)=cos(t)$
- 正态分布，概率密度函数 $f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$,$\phi(t)=e^{-\frac{t^2}{2}}$
- 泊松分布 $p(n)=e^{-\lambda}\frac{\lambda^n}{n!}$,$\phi(t)=e^{-\lambda(1-e^{it})}$

复习一个重要极限

自然对数底数 e 的定义

- $\lim_{n\to\infty}(1+1/n)^n$ 存在，且定义 $e=\lim_{n\to\infty}(1+1/n)^n$
- 于是定义 $e^x=\lim_{n\to\infty}(1+x/n)^n$

大数定律和中心极限定理

大数定律

X 是随机变量，μ 是 X 的期望，σ 是 X 的方差.$\{X_k\}_{k=1}^{\infty}$ 是服从 X 的独立同分步随机变量，那么 $\bar{X}_n = \frac{\sum\limits_{k=1}^{n} X_k}{n}$ 依概率收敛于 μ. 也就是说对于任何 $\epsilon > 0$ 有

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

Proof

因为 X 具有一阶矩，所以特征函数 $\phi_X(t)$ 存在一阶泰勒展开 $\phi_X(t) = 1 + i\mu t + o(t)$，于是

$$\begin{aligned}
\phi_{\overline{X}}(t) &= E(exp(it\frac{\sum_{i=1}^{n}}{n})) \\
&= \prod_{i=1}^{n} E(exp(itX/n)) \\
&= (1 + \mu t/n + o(t/n))^n
\end{aligned}$$

Continue Proof

于是

$$\lim_{n\to\infty}\phi_{\overline{X}}(t) = \lim_{n\to\infty}(1+i\mu t/n+o(t/n))^n$$
$$= e^{i\mu t}$$

这就是独点分布的特征函数，所以 $\overline{X}$ 按分布收敛于独点分布.

按分布收敛于独点分布就等于按概率收敛于一个常数.

中心极限定理

X 是随机变量,$\phi(X)$ 是 X 的特征函数.$\{X_k\}_{k=1}^{\infty}$ 是服从 X 的独立同分步随机变量，那么

$$Z_n = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$$

依分布收敛于正态分布 $N(0,1)$.
也就是说对于任何 $\epsilon > 0$ 有

$$\lim_{n\to\infty} P(Z_n < z) = \Phi(z), \quad \forall z$$

其中 Φ 是标准正态分布的分布函数.

Proof

因为 X 具有一阶矩和二阶矩，所以特征函数 $\phi_X(t)$ 存在二阶泰勒展开

$$\phi_X(t) = 1 + i\mu t - \frac{\sigma^2}{2}t^2 + o(t^2)$$

令 $Y = (X-\mu)/\sigma$, 则 $E(Y) = 0, E(Y^2) = 1$, 于是有

$$\phi_Y(t) = 1 - \frac{1}{2}t^2 + o(t^2)$$

因为 $Z_n = \sqrt{n}\overline{Y}$，所以

$$\begin{aligned}\phi_{Z_n}(t) &= E(exp(it\sum_{i=1}^{n} Y_i\sqrt{n})) \\ &= (1 - \frac{1}{2n}t^2 + o(t^2/n))^n\end{aligned}$$

Continue Proof

于是

$$\lim_{n\to\infty}\phi_{Z_n}(t) = \lim_{n\to\infty}(1-\frac{t^2}{2n}+o(t^2/n))^n$$
$$= e^{-\frac{1}{2}t^2}$$

这就是正态分布的特征函数，所以 Z_n 按分布收敛于正态分布.

小结 (切比雪夫不等式，大数定律，中心极限定理)

- 随机变量的矩可以描述随机变量所服从分布的性质
- 随机变量的特征函数可以全面描述随机变量的分布
- 切比雪夫不等式指出方差可以描述随机变量取值的分散程度
- 大数定律指出独立重复实验的平均值的收敛规律
- 中心极限定理给出独立重复实验平均值更细致的描述

谢谢大家!